Improving Translation Quality of Rule-based Machine Translation

Paisarn Charoenpornsawat

Virach Sornlertlamvanich

Thatsanee Charoenporn

National Electronics and Computer Technology Center
THAILAND

Agenda

- ✶ Introduction.
  - – MT approaches, Why we improve RBMT?,
- ✶ A rule-based machine translation approach.
- ✶ Applying machine learning technique.
- ✶ An overview of the system.
- ✶ Preliminary experiments & results.
- ✶ Conclusion.

Introduction

- MT has been developed for many decades.
- Many approaches have been proposed such as rule based, statistic-based and example-based approaches.
- No approach produces a translation quality that meets human’s requirements.
- Each approach has its own advantages and disadvantages.

Machine Translation Approaches.

- A rule-based approach.
  - It can deeply analyzes in both syntax and semantic levels.
  - It uses much linguistic knowledge.
  - It is impossible to write rules cover the whole of a language.
  - The translation accuracy depends on linguistic rules.
- A statistic-based approach.
  - It does not require linguistic knowledge.
  - It needs statistics of bilingual corpus and a language model.

Machine Translation Approaches. (cont.)

  - It can produce a suitable translation even if a given sentence is not similar to any sentences in the training corpus.
  - It can not translate idioms and phrases that reflects long-distance dependency.

- An example-based approach.
  - It does not require linguistic knowledge.
  - It uses large bilingual corpus.
  - It can only produce suitable translations in case of a given sentence must similar to any sentences in the training data.

Why we decided to improve a Rule-based Machine Translation ?

- Most of commercial MT products in market are using rule-based approaches.
- A statistic-based and example-based approaches are need large bilingual corpus.
- Rules in RBMT are produced from linguistic knowledge.
- RBMT can deeply analyze in both syntax and semantic levels. So it can give syntax and semantic information.

Case Study In a rule-based machine translation.

ParSit: Eng-Thai MT.

- *ParSit* is an English to Thai machine translation that provides a free service on www.suparsit.com.
- It is an interlingual-based approach.
- *ParSit* consists of four modules.

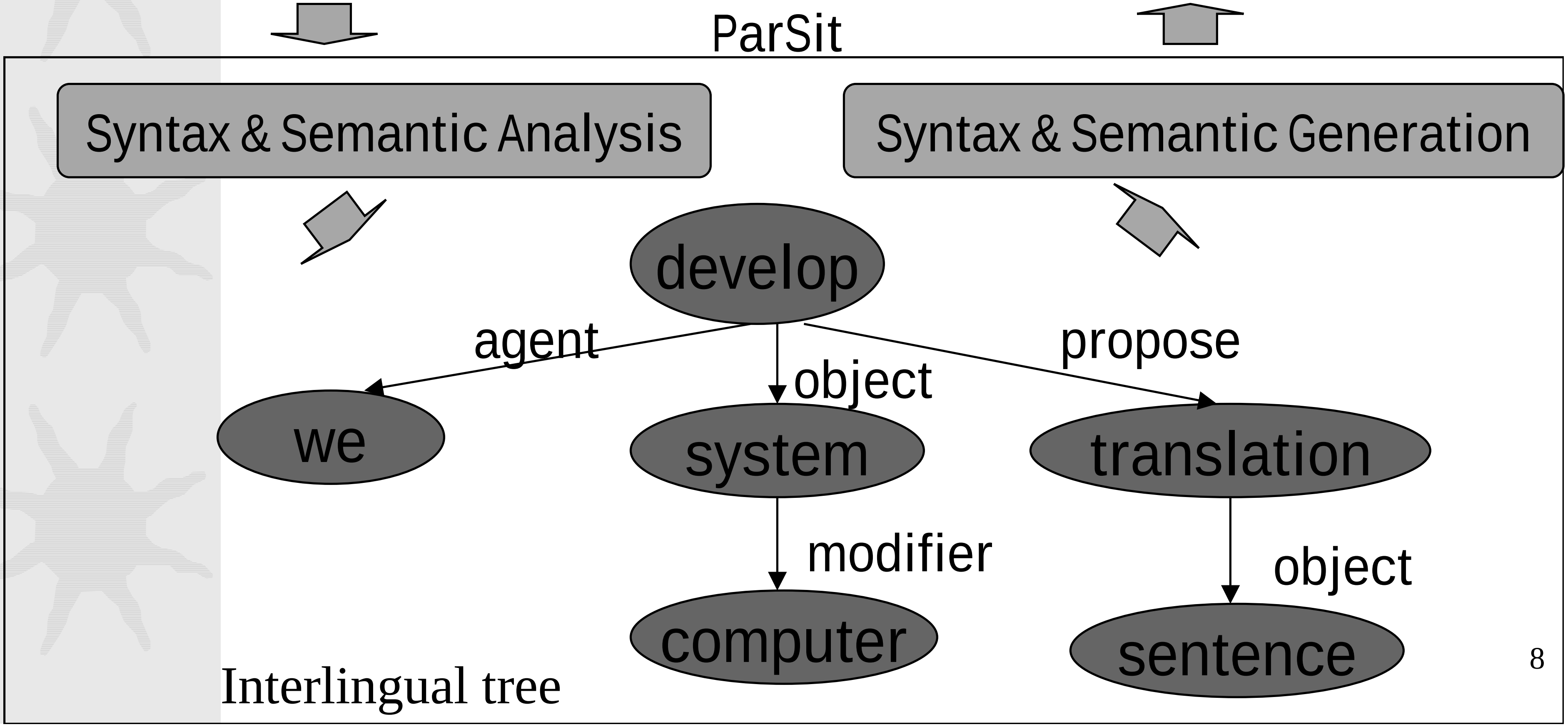
  1.) Syntax analysis *2.) Semantic analysis*

  3.) Syntax generation *4.) Semantic generation*

ParSit Translation Process.

We develop a computer system
for sentence translation.

พวกเรา พัฒนา ระบบ คอมพิวเตอร์ เพื่อ
การแปล ประโยค



ParSit

Syntax & Semantic Analysis

Syntax & Semantic Generation

develop

agent

object

propose

we

system

translation

modifier

object

computer

sentence

Interlingual tree

Errors of translation

- ✶ We classify an error of translation into two main groups.
  1. Incorrect meaning errors.
  2. Incorrect ordering errors.
- ✶ Incorrect meaning errors can be divided into 3 subgroups.
  – Missing some words.

<u>The city is not far from here</u>

เมือง ไม่ ไกล จาก นี่ incorrect

เมือง <u>อยู่</u> ไม่ ไกล จาก นี่ correct

Errors of translation (2)

– Generating over words.

This is the house in which she lives.

นี่ คือ บ้าน ที่ เธอ อาศัย อยู่ ~~ที่ ในนั้น~~ incorrect

นี่ คือ บ้าน ที่ เธอ อาศัย อยู่ correct

– Using an incorrect word.

The news that she died was a great shock.

ข่าว ที่ว่า ที่ เธอ ตาย เป็น ที่ตกใจ ที่ยิ่งใหญ่ incorrect

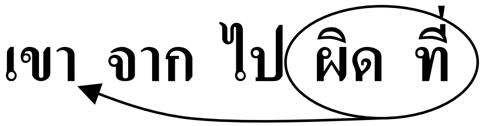
↑

ข่าว ที่ว่า ที่ เธอ ตาย เป็น ที่ตกใจ อย่างมาก correct

Errors of translation(3)

✶ *Incorrect ordering errors*.

He is wrong to leave.



เขา จาก ไป ผิด ที่ incorrect

เขา ผิด ที่ จาก ไป correct

Statistics of *ParSit* Errors

Incorrect meaning errors			Incorrect ordering errors (%)
M (%)	G (%)	U (%)	
16.71	13.31	51.42	18.26

The traditional method in improving a RBMT

- To improve quality of a RBMT, we have to modify rules.
- This method requires much linguistic knowledge.
- It cannot guarantee that the overall accuracy will be better.

Concepts of our system

- ✶ The main problems of translation are choosing incorrect meaning.
- ✶ It can be view as a classification or disambiguation problem
- ✶ To improve the accuracy, we apply a method to disambiguate meanings of only a word in question.
- ✶ The context of a word in question will use in disambiguation.

Why we apply ML techniques to RBMT?

- A ML technique is an adaptive model.
- It do not need linguistic knowledge.
- It can automatically extract useful information from the training data.
- Many ML techniques highly success in classifying problems.

Machine Learning Techniques

- ✶ Machine learning techniques automatically extract the context features that useful information in disambiguating a word in question.
- ✶ C4.5, C4.5rule and RIPPER were selected in our experiment.

C4.5 & C4.5rule

- C4.5, decision tree, is a traditional classifying technique that proposed by Quinlan (1993).
- C4.5rule is extended from C4.5. It extracts production rules from an unpruned decision tree produced by C4.5, and then improves process by greedily deletes or adds single rules in an effort to reduce description length.

RIPPER

- RIPPER is a propositional rule learning algorithm that constructs a ruleset which classifies the training data.

Ruleset :

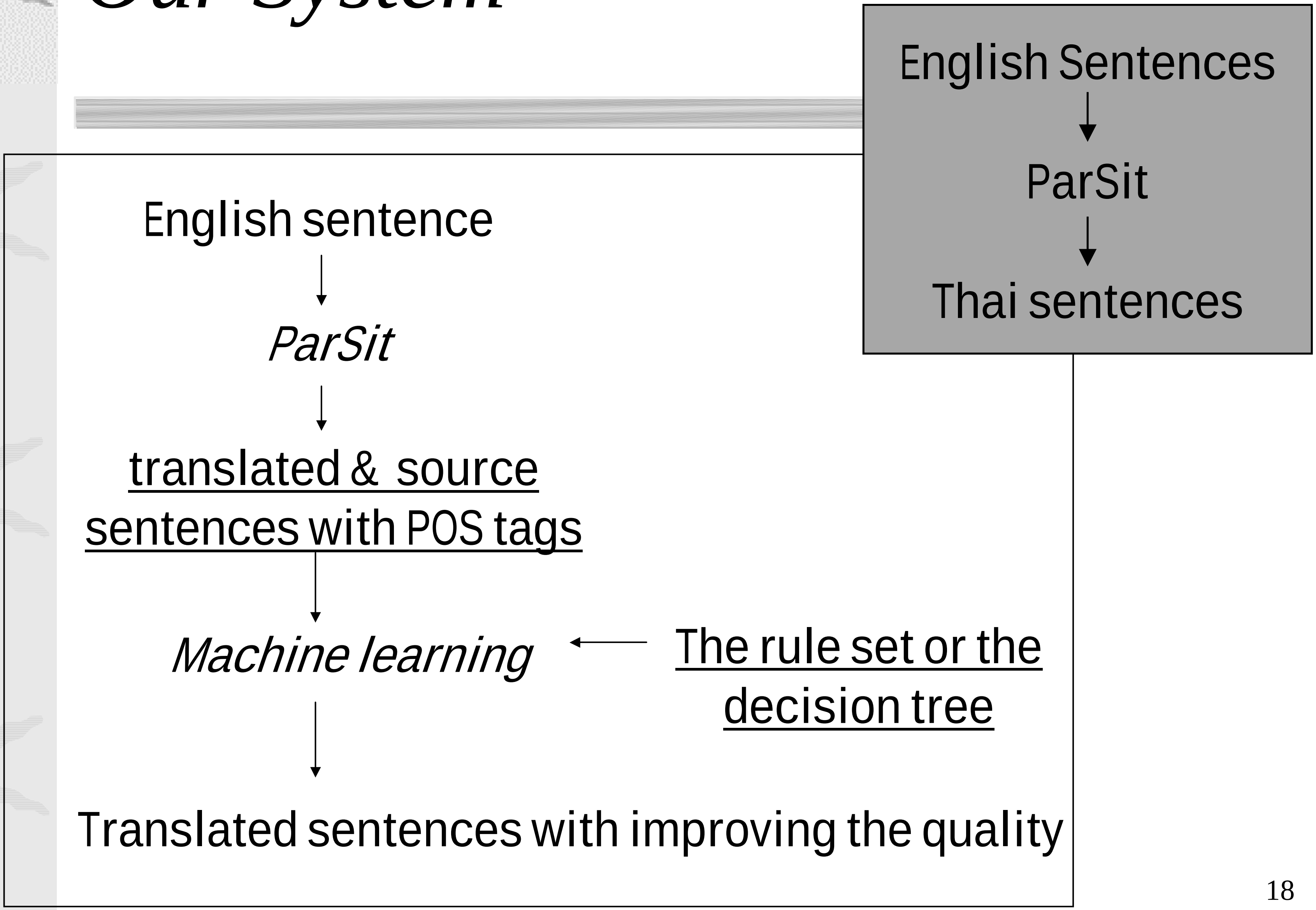
if T_1 and T_2 and … T_n then class C_x

T_i is a condition.

C_x is the target class to be learned.

Our System

Normal translation



English Sentences

ParSit

Thai sentences

English sentence

ParSit

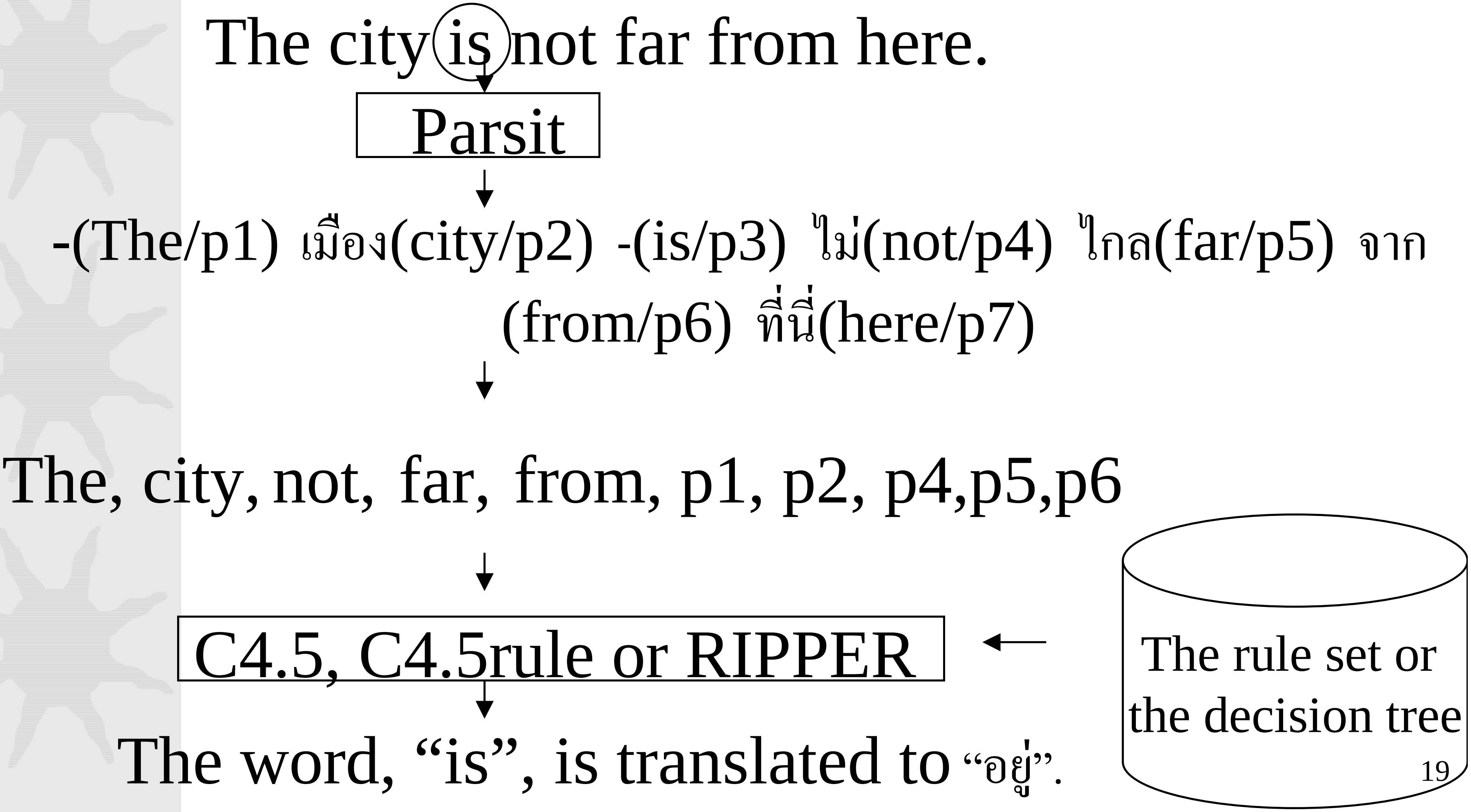
translated & source sentences with POS tags

Machine learning

The rule set or the decision tree

Translated sentences with improving the quality

An example of translation



The city is not far from here.

Parsit

-(The/p1) เมือง(city/p2) -(is/p3) ไม่(not/p4) ไกล(far/p5) จาก (from/p6) ที่นี่(here/p7)

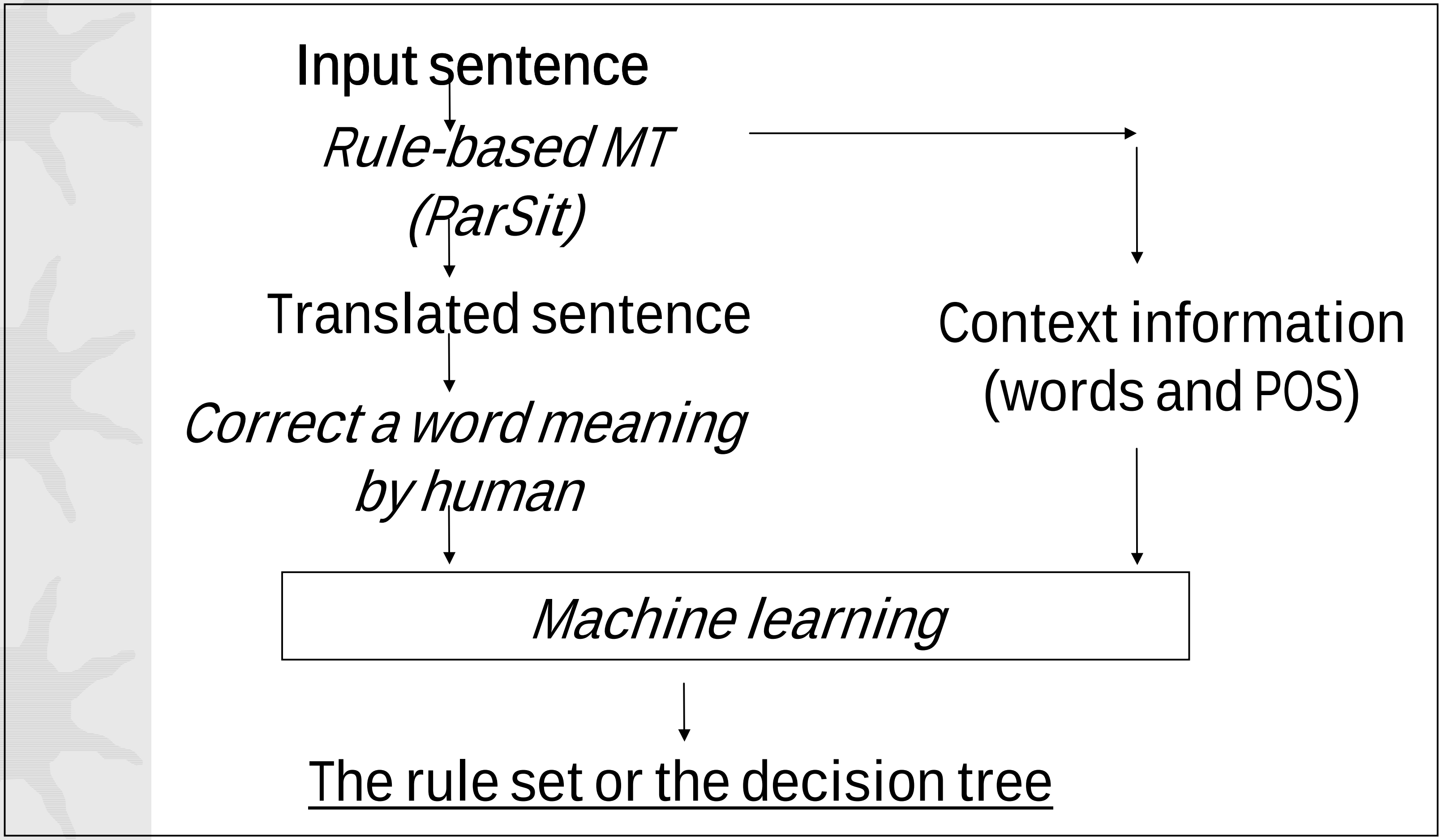
The, city, not, far, from, p1, p2, p4,p5,p6

C4.5, C4.5rule or RIPPER

The rule set or the decision tree

The word, "is", is translated to "อยู่".

Our System (2) The training module



Input sentence

Rule-based MT (ParSit)

Translated sentence

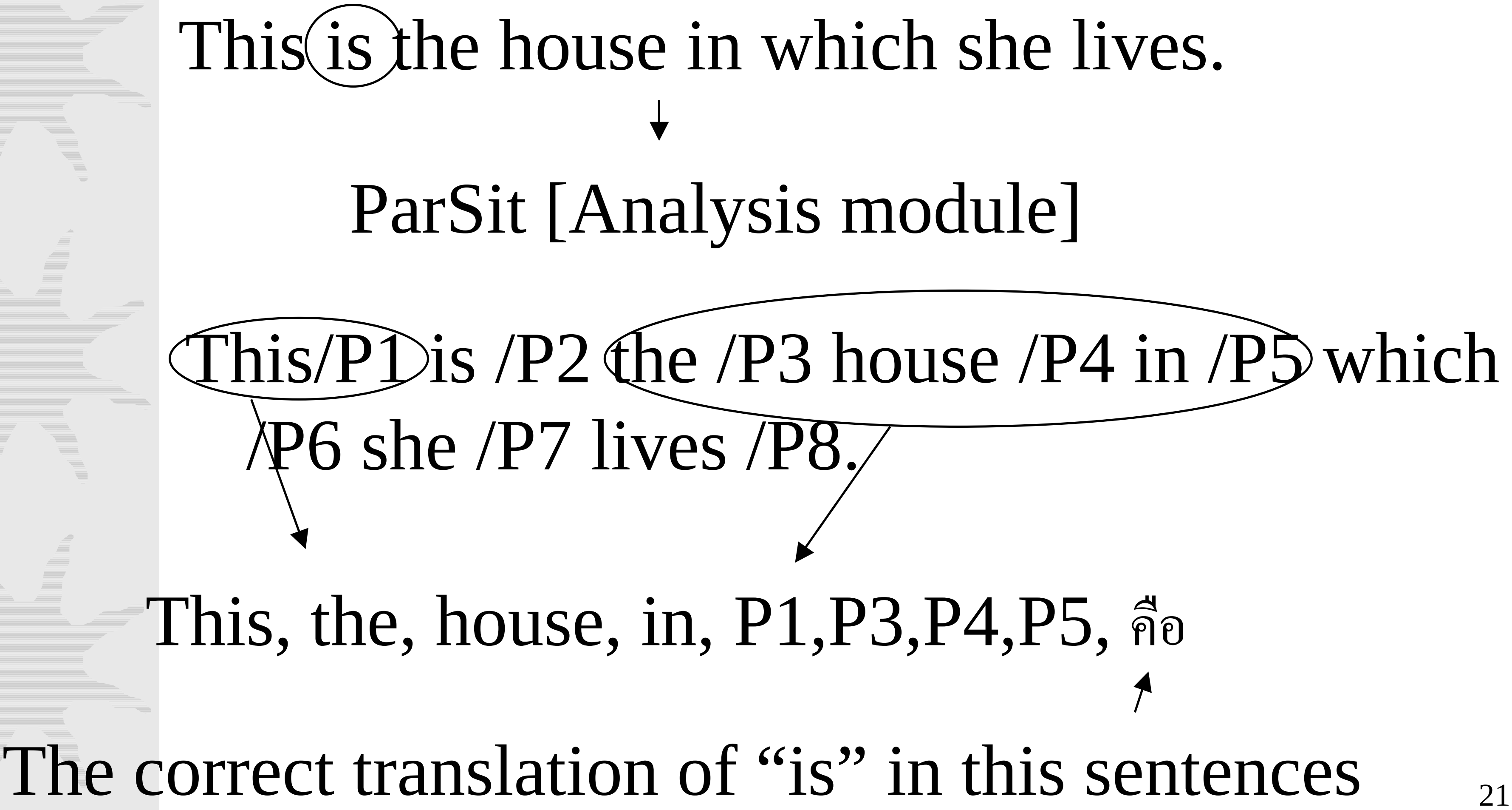
Correct a word meaning by human

Context information (words and POS)

Machine learning

<u>The rule set or the decision tree</u>

An example of training data



This is the house in which she lives.

ParSit [Analysis module]

This/P1 is /P2 the /P3 house /P4 in /P5 which /P6 she /P7 lives /P8.

This, the, house, in, P1,P3,P4,P5, คือ

The correct translation of “is” in this sentences

Preliminary Experiments

- ✶ An verb-to-be is the first target for testing because it frequently appeared.
- ✶ It quite difficult in translation into Thai by using only linguistic rules. (48% accuracy by ParSit)
- ✶ 3,200 English sentences from EDR corpus were selected in our experiments.
- ✶ We used 700 sentences for testing and the rest for training.
- ✶ We tested on different sizes of training data and features.

Results

The results from C4.5

	100	500	1 K	1.5K	2K	2.5K
Pos±1	**67.1**	**69.8**	69.8	69.8	69.8	69.8
Pos±2	**67.1**	**69.8**	69.8	69.8	69.8	69.8
Pos±3	**67.1**	**69.8**	69.8	69.8	69.8	69.8
Word±1	55.5	63.2	**73.1**	**74.2**	**75.5**	75.4
Word±2	57.7	64.6	71.7	72.7	**75.5**	77.3
Word±3	57.8	65.3	71.3	73.1	75.4	**77.7**
P&W±1	55.5	68.6	71.1	71.3	71.8	71.8
P&W±2	57.7	68.6	71.3	70.4	71.8	71.8
P&W±3	57.8	68.6	71.3	69.6	71.3	71.9

Results (2)

The results from C4.5rule

	100	500	1 K	1.5K	2K	2.5K
Pos±1	**69.8**	71.3	76.3	**77.3**	76.0	**73.1**
Pos±2	**69.8**	**77.5**	**76.7**	76.9	76.3	**73.1**
Pos±3	69.2	77.2	76.2	76.8	70.1	**73.1**
Word±1	54.9	73.1	63.4	63.6	67.2	71.1
Word±2	56.3	73.5	73.5	72.5	64.7	70.6
Word±3	56.3	72.2	72.5	72.3	**76.8**	70.6
P&W±1	54.9	77.2	63.4	68.4	69.2	71.1
P&W±2	56.8	76.7	73.5	68.0	70.5	70.6
P&W±3	56.8	69.6	64.3	61.8	71.5	71.1

Results (3)

The results from RIPPER

	100	500	1 K	1.5K	2K	2.5K
Pos±1	70.2	70.9	**73.3**	71.7	72.1	**76.1**
Pos±2	69.4	71.0	69.2	70.2	70.8	72.1
Pos±3	69.2	71.0	69.6	71.3	**76.9**	70.6
Word±1	63.1	69.8	67.2	72.1	72.9	71.1
Word±2	55.3	67.7	66.8	**74.0**	72.2	70.6
Word±3	58.0	70.5	66.8	71.7	72.3	70.6
P&W±1	**72.7**	**73.9**	**73.3**	73.5	73.4	**76.1**
P&W±2	57.7	72.3	69.2	73.5	72.2	72.1
P&W±3	62.0	70.4	69.6	72.1	72.6	70.6

Conclusion

- C4.5, C4.5rule and RIPPER have efficiency in extracting context information from a training corpus.
- The accuracies of these three ML techniques are not quite different.(about 77% accuracy)
- RIPPER gives the better results than C4.5 and C4.5rule in a small train set.
- The best feature for our problem depending on the a machine learning technique.

Conclusion (2)

- The suitable context information giving the highest accuracy in C4.5, C4.5rule and RIPPER are ±3 words, ±2 POS tags and ±1 word & POS tags respectively
- Our idea can be apply to any RBMT and it do not require bilingual corpus.
- In future, we will increase the data size, features and words in question.

Thank you